SEMANTIC INDEXING FOR COMPLEX PATIENT GROUPING

Kilian STOFFEL[†‡] Joel SALTZ[†‡], Jim HENDLER[†], Jim DICK[‡], William MERZ[‡] and Robert MILLER[‡]

[†]Computer Science Department
University of Maryland
College Park, 20742, MD

[‡]Johns Hopkins Hospital
Department of Pathology
Baltimore, 20885, MD

Introduction. A joint effort between computer scientists at the University of Maryland and clinicians at Johns Hopkins Hospital focuses on using high performance computing technology in support of medical applications. We describe a project that focuses on providing database support to microbiologists and infectious disease specialists.

The taxonomy used to describe micro-biological data is complex; this terminology also changes with time. A problem we have encountered is the difficulty clinicians and researchers face in formulating queries that capture the kinds of questions they wish to pose. The result of this project will be to provide medical specialists with the ability to express complex queries without becoming experts on the underlying data model.

As a means for overcoming this sort of difficulty, we are looking at how to efficiently integrate "semantic" knowledge, stored in the form of thesauri (or more technically, ontologies) with low-level data to allow efficient indexing of large databases. We create a mapping between the end-user terminology and attribute-value pairs in a relational database. We optimize performance by indexing the whole database using an ontology that represents the end-user's terminology

Motivating Examples. The complex taxonomy used to describe microorganisms and the non-uniform degree of microorganism identification create difficulties for those who wish to pose queries to a clinical data warehouse. We will provide a brief description of the functionality provided by our system in the context of an example of a specific query that is of clinical interest at Johns Hopkins Hospital.

The clinical context is an ongoing effort to characterize microorganism antibiotic susceptibility. This issue is of continuing concern because of the emergence of antibiotic resistant microorganisms. The query can be informally stated as follows:

Which antibiotics are effective against 80% of organisms recovered from cultures that grew out any species of Enterobacteriaceae?

Our tools support queries that characterize the effectiveness of antibiotics with respect to many possible categories of organisms. The two categorical terms (antibiotics and Enterobacteriaceae) allow us to discuss two types of categories. The first term "Antibiotics" is replaced by a list of all antibiotics tested. The second term "Enterobacteriaceae" is replaced by all its sub-categories. Sub-categories of Enterobacteriaceae are not disjoint. For instance, *Citrobacter* constitutes one subcategory and *Citrobacter diversus* and *Citrobacter freundii* constitute two additional sub-categories.

In our database, we maintained tests for 27 antibiotics and we store 65 sub-concepts of Enterobacteriaceae. Thus Example 2 allows us to evaluate antibiotic susceptibilities for 1755 different combinations of organism and antibiotic. This query required roughly 40 minutes on a 150 MHz Pentium PC using a database with 20521 records of organisms.

The scope of this abstract precludes a detailed discussion of results obtained. We found, for instance, that over 80% of our Enterobacteriaceae were susceptible or very susceptible to amikacin, ciprofloxacin, ceftazidime, cefuroxime, Gentamicin, piperacillin, trimethoprim/sulfa, ticarcillin, and tobramycin. For all species of *Citrobacter* taken together, only amikacin, ciprofloxacin, gentamicin, and tobramycin were effective against 80% of specimens. However, for *Citrobacter diversus* three additional antibiotics were effective in 80% of specimens - ceftazidime, cefuroxime and trimethoprim/sulfa.

Conclusion and Future Work. A prototype of the system is currently being tested in the Johns Hopkins Hospital. In the longer term, we will use these algorithms as the base for a more general decision support system. Furthermore, we are currently exploring the use of this indexing technique in other domains in which complex data retrieval is used and where ontologies can be generated.

1091—8280/97/$5.00 © 1997 AMIA, Inc.